WORKING PAPER 61

MATHEMATICA Policy Research

Working PAPER

BY DANIEL KASSLER, IRA NICHOLS-BARRER, MARIEL FINUCANE

Beyond "Treatment versus Control": How Bayesian Analysis Makes Factorial Experiments Feasible in Education Research

June 2018

ABSTRACT

Researchers often wish to test a large set of related interventions or approaches to implementation. A factorial experiment accomplishes this by examining not only basic treatment-control comparisons but also the effects of multiple implementation factors such as different dosages or implementation strategies, and the interactions between these factor levels. However, traditional methods of statistical inference may require prohibitively large sample sizes to perform complex factorial experiments.

In this paper, we present a Bayesian approach to factorial design. By using hierarchical priors and partial pooling, we show how Bayesian analysis substantially increases the precision of estimates in complex experiments with many factors and factor levels, while controlling the risk of false positives from multiple comparisons. Using an experiment we performed for the U.S. Department of Education as a motivating example, we perform power calculations for both classical and Bayesian methods. We repeatedly simulate factorial experiments with a variety of sample sizes and numbers of treatment arms to estimate the minimum detectable effect (MDE) for each combination.

We found that the Bayesian approach yields substantially lower MDEs when compared with classical methods for complex factorial experiments. For example, to test 72 treatment arms (five factors with two or three levels each), a classical experiment requires nearly twice the sample size as a Bayesian experiment to obtain a given MDE. Overall, our study showed that Bayesian methods are a valuable tool for researchers interested in studying complex interventions. They make factorial experiments with many treatment arms vastly more feasible.

I. BACKGROUND AND MOTIVATION

Most social policy experiments compare a single treatment group with a control group. For example, a typical study in the field of education research might compare the effect of introducing a new set of mathematics lesson plans with that of maintaining the status quo. However, this type of study is limited to a small number of treatment arms—typically only one or two approaches are tested. A variety of open research questions in education would benefit from a design that makes it possible to evaluate many different practices in one study, as there are often a number of interventions of interest in a given topic area, and (within each of those) there are many different ways to implement a given program.

Factorial experiments, which have long been used in agriculture and engineering (Cox 1958), allow researchers to efficiently test a larger and richer set of related programs or practices in a single study. A factorial experiment moves beyond basic treatment-control comparison and examines the effects of multiple implementation factors such as different dosages and implementation strategies, along with the interactions between these factor levels. For example, in a hypothetical study of new math curricula, a researcher might want to study different lesson lengths (30, 60, or 90 minutes per day) and different teaching strategies (whether to deliver lessons using standard classroom teachers or math specialists). Because there is a possibility of interaction effects (math specialists may be particularly effective in a longer teaching session), these two factors cannot be tested separately. A factorial experiment assigns classes to a random lesson length and random teaching strategy, so that all combinations of lesson length and teaching strategy are tested. This design has two factors (lesson length and teaching strategy) that have three and two levels respectively (30, 60, and 90 minute lessons; standard teachers and math specialists). Table 1 shows the six treatment arms of this example (3 × 2) configuration. In addition to reflecting differences in treatment, the factors of a factorial experiment can also encode differences in treated populations (such as by age groups or gender). Such an experimental design may be useful for studying heterogeneity in treatment effects by identifying the groups for which a treatment is most or least beneficial.

Table 1. Illustrative framework for a 3 × 2 study design

		Factor B: Teaching strategy	
		Standard teacher	**Math specialist**
Factor A: Lesson length	**30 minutes**	Standard teacher for 30 minutes	Math specialist for 30 minutes
	60 minutes	Standard teacher for 60 minutes	Math specialist for 60 minutes
	90 minutes	Standard teacher for 90 minutes	Math specialist for 90 minutes

Note: This table illustrates the structure of a simple factorial experiment on math classes. The experiment has two factors (lesson length and teaching strategy) that have three and two levels respectively in a (3 × 2) design. Each cell of the table indicates a unique combination of factor levels that defines one of the study's treatment arms.

Historically, factorial experiments in education have been rare, in part because of the large sample sizes required. Such large sample sizes often come with major cost and logistical challenges, as compliance with treatment must be maintained across multiple treatment arms; in

addition, managing compliance with large samples can be particularly complex in an education setting. There are also methodological challenges associated with large sample sizes. In a conventional factorial design seeking to estimate the effect of a number of treatment arms, each treatment arm requires its own independent hypothesis test. This means an experiment with many treatment arms will have to run many such independent tests. As the number of hypothesis tests increases, so does the probability that at least one of them will yield a false positive—a situation referred to as the multiple comparisons problem (Waller and Duncan 1969). The large number of contrasts in a factorial experiment makes it especially susceptible to this issue. The U.S. Department of Education's Institute for Education Sciences has been particularly focused on this problem. Specifically, it has established strict guidelines for accounting for multiple comparisons in the What Works Clearinghouse standards, which are used to assess the results of impact studies throughout the education field. Although correcting for multiple comparisons is possible, the most common ways of doing so effectively apply a post-hoc penalty on the precision of the experiment, decreasing the risk of false positives at the cost of increasing the likelihood of false negatives (Schochet 2008). Larger sample sizes lead to higher precision, mitigating the downside of these post-hoc corrections, but acquiring a sample size large enough to test more than a few treatment arms is often challenging.

Due to sample size constraints, many guidelines for conducting factorial experiments in the realms of social policy recommend approaches that either limit the number of tested factors or recommend the use of fractional factorial designs that selectively omit certain factor combinations from the experiment (for example, Chakraborty et al. 2009; Collins et al. 2009; Dziak et al. 2012; Nair et al. 2008). These approaches can be extremely useful when the research questions a study is seeking to answer are well defined and relatively small in number. If a study is only seeking to test a limited number of factors (and there are only a few tested levels within each factor), sample size constraints are less of an issue. Similarly, if a study is able to ignore some or all potential interaction effects between factors (that is, if there is a strong theoretical basis to believe that factors are simply additive in their effects), fractional factorial designs can substantially reduce sample size requirements as well. However, when researchers seek to investigate a large number of factors and account for interaction effects between factors, these sample-size constraints are more likely to be prohibitive.

Bayesian inference, which uses data from the experiment alongside any available prior information about model parameters and the relationships among them, makes it possible to overcome these challenges and efficiently conduct large factorial studies. In particular, large factorial experiments benefit from analysis with hierarchical Bayesian models. In such analyses, impact parameters of interest themselves are treated as random effects (an approach often associated with the work of Andrew Gelman); older methods that treat these parameters as fixed would lack these advantages. In this paper, we describe a Bayesian framework for factorial experiments and illustrate its use with an example drawn from our work in the field of education research. We begin by introducing the set of research questions that motivated us to design a factorial experiment. We then introduce the Bayesian approach to analyzing results from a factorial experiment, both in general and as it was implemented in our study. We demonstrate the utility of this approach with a power analysis and examine the circumstances under which the design is likely to prove most useful to other researchers.

II. AN EXAMPLE: CREATING SCHOOL PROFILES FOR PARENTS

We developed a factorial study design as part of a forthcoming U.S. Department of Education study that tests methods for presenting school choice information to low-income parents (Nichols-Barrer et al. 2016).[1] The study investigated how best to help parents make informed decisions about which school to select for their children, in the context of open enrollment policies that enable parents to choose between large numbers of public schools (such as magnet or charter schools). Specifically, our study aimed to generate evidence about how to improve school choice information displays or guides. These displays (often created by school district or state education officials) usually present a set of school profiles to parents, with information about each school that is intended to help parents make informed school selections. They are typically published as websites; a common format is to include a map of the school district along with a list of short profiles or school website links. The study was designed to answer a broad set of research questions pertaining to the effects of different displays on users. In particular, we examined the effects of different display strategies on the following three concepts:

- Understandability (ability of users to process factual information)
- Perceived usability (how easy or satisfying parents find a display to use)
- Actual school selections (which schools each parent is likely to pick for his or her child)

A factorial design was the natural choice for several reasons. First, there are a large number of independent decisions to make when creating a school information display. Examples include the amount and type of information to show for each school, the format in which to present performance indicators, and the default sort ordering of the schools shown in the display. Second, interactions are very likely to occur between these display elements. For example, the effect of adding parent ratings to a display may be beneficial when the overall amount of information is low, but when the display is already complex the addition of another data source could be overwhelming. An experiment that only tested a small number of displays would obscure these types of interactions. To identify which of the many differences between any two displays was responsible for a given impact on parents, the experiment needed to test all combinations of potential display choices. A factorial experiment provided an appropriate framework to match these needs.

The experiment started with a basic website template that remained the same across all treatment arms, and consisted of a map showing school locations at the top followed by a list of schools. Four categories of information were shown for each school: distance to school, academic performance, safety, and school resources. Based on the results of our power calculations, we tested a total of five factors in a ($3 \times 3 \times 2 \times 2 \times 2$) configuration, for a total of 72 distinct treatment arms. In this study, the five factors were not examined independently: rather, the experiment sought to identify which of the 72 possible combinations of factor levels

[1] We are members of a larger research team who conducted the study for the U.S. Department of Education. We designed the study's analytical approach and carried out data analysis, while a broader team developed the list of tested factors, created the information displays, and managed survey operations. The overall study was directed by Steve Glazerman; the information displays used in the experiment were created by Tembo, Inc.

represented the best possible design of an information display, after accounting for the interaction effects between factors. In other words, the study sought to identify which of the 72 treatment arms represented the best possible display for each outcome.

Each factor reflected a type of display feature that varied within the template. The five factors were (1) amount of information (the number of indicators in each information category); (2) information format (numbers, graphs, or letter-grade icons); (3) use of a reference point (inclusion or absence of district averages); (4) data sources (inclusion or absence of parent ratings); and (5) default sort order (by distance to school or by academic performance). Table 2 shows the full list of factors used, along with their levels. Figure 1 illustrates the experiment with an example of an information display shown to parents and diagrams how these variations affected it.

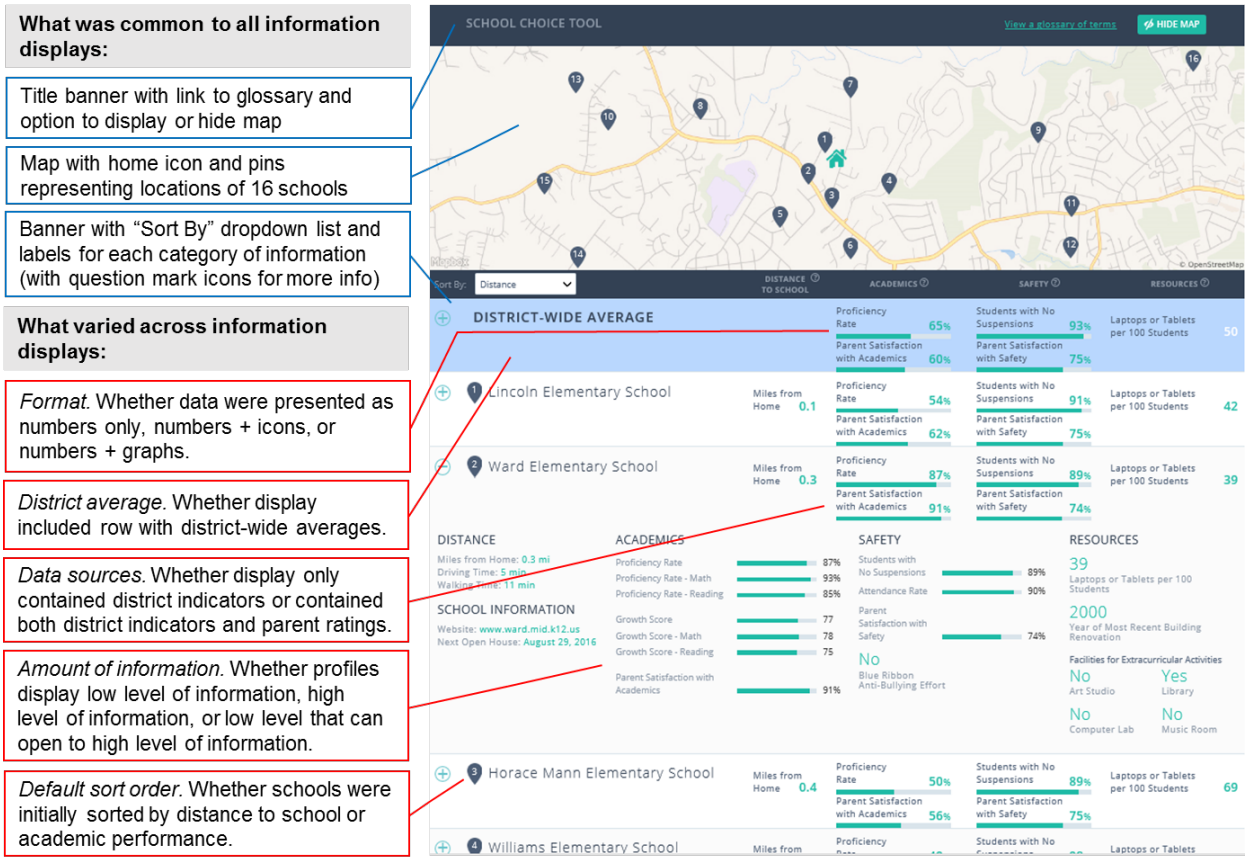
Table 2. Factors and factor levels in the experiment

Factor	Level 1	Level 2	Level 3
A. Format	Numbers	Numbers + icons	Numbers + graphs
B. District average	No district average	District average shown	n.a.
C. Data sources	District only	District + parent ratings	n.a.
D. Amount of information	Lower amount: one attribute per domain	Higher amount: multiple attributes per domain all shown at once	Progressive disclosure: lower information by default, with option to expand the view to the higher amount
E. Default sort order	By distance	By academics	n.a.

Note: This table shows all levels of the five factors used in our study.

n.a. = not applicable.

Figure 1. School information displays tested in the experiment



Note: This figure illustrates one of the 72 information displays prepared for the experiment. On the right of the screen is a school display with the following factor levels: drop-down information, school data formatted as graphs, inclusion of a district reference point, inclusion of parent ratings, and default sort by distance to the school. The left side of the figure indicates where and how each factor modifies the display.

The study was designed to test whether different display designs can influence the types of schools parents select for their children (for example, whether initially sorting the schools by academic quality can "nudge" parents to select higher-quality schools). The study also tested whether these information displays had an effect on the outcomes of understandability (whether parents correctly answered factual questions about schools in the display) and usability (whether parents reported that the display was easy to use or satisfying). One benefit of Bayesian methods is that posterior distributions permit the study to make direct probabilistic statements about which display features are best. As such, the analysis reported the probability that each display strategy outperformed the other tested levels of each factor and identified which combination of strategies was best for each outcome.

III. THE BAYESIAN FACTORIAL DESIGN

In addition to the data from the experiment, Bayesian analysis requires the researcher to set pre-specified probability distributions for the analytic model parameters to help fit the model. These prior distributions (or simply priors) allow the experimenter to incorporate previously

known information about (1) the values of the parameters or (2) the relationships among parameters in the model. We will discuss each of these purposes in turn.

A researcher unfamiliar with Bayesian inference may be concerned that introducing outside information about parameter values would allow experimenters to bias results by incorporating their previous expectations into the model. Although this is possible in theory, these concerns are manageable in practice. When used appropriately, priors are based on reasonable expectations about the parameters, often from reviews of related literature. For example, an education researcher might use a standard normal distribution as a prior if other experiments in education rarely show effect sizes larger than 1 standard deviation, but should take care to center this prior at a mean of zero to remain agnostic about whether the treatment will be successful. Priors also become less important as sample sizes increase. In an experiment with adequate sample size,[2] the conclusions drawn about parameter values will largely come from the data in the experiment, and not the prior (Gelman et al. 2015). (Even though we argue that Bayesian methods provide gains in precision compared to classical methods, designers of Bayesian experiments must still use power calculations to assess the adequacy of a given sample size.)

Prior distributions can also reflect information about the relationships between parameters of interest, which has important consequences for a factorial experiment. In particular, using so-called hierarchical Bayesian prior distributions (to reflect the hierarchy of levels nested within factors within the experiment as a whole) has two chief benefits. The first is to achieve what is known as partial pooling, which can improve the statistical precision of effect size estimates profoundly. Partial pooling refers to the fact that a Bayesian approach can model the effects of each level of a given factor with a single shared prior distribution for that factor. For example, in our study we model the effects of sorting by distance and sorting by academics as coming from a shared prior distribution (and, likewise, each other factor has a prior distribution shared across its levels). By inducing partial pooling, the researcher supposes that the distribution of effect sizes within the same factor (for example, sort order) may have a distinct variance from the distributions found in other factors in the experiment (format, amount of information, and so on). This allows data from the experiment to identify which factors are most important (in the sense that the variance of effect sizes within those factors are larger than the variance of effect sizes within other factors) (Gelman 2005). The term partial pooling refers to the process of pooling observations together across all levels of a factor when estimating the effect of each of the factor's levels, especially when that factor proves to be relatively unimportant compared with other factors. Within a given factor, the result is that the estimates for the effects of each level are informed by one another, leading to larger effective sample sizes and smaller uncertainty in estimates. The variance parameters of these priors are also partially pooled to borrow

[2] Although priors may be more determinative of final estimates for studies with very small sample sizes, such studies are likely to experience many other problems as well. Indeed, the importance of adequately powered studies is often underestimated. It is tempting to interpret significant findings in an underpowered experiment as especially notable, having achieved the requisite threshold of significance despite the study's low power; however, the effect estimates are often of substantially overestimated magnitude (sometimes by more than tenfold) and have a high probability (in extreme cases, nearly 50 percent) of being the wrong sign. See Gelman and Carlin (2014) for more details on these phenomena.

information about the overall effect size across factors, providing greater stability in estimates of factors with few levels (Gelman and Hill 2007).[3]

The second advantage of using a hierarchical Bayesian approach to reflect information about the structure of the model is to account for multiple comparisons. Classical statistical procedures typically perform many hypothesis tests and then correct for the problem of multiple comparisons by inflating confidence interval widths or decreasing the p-value cutoff for statistical significance, without adjusting effect estimates themselves (Benjamini and Hochberg 1995). Although this reduces the risk of incorrectly identifying effects as significant (reducing Type I error), it does so at the cost of obfuscating potentially important effects (increasing Type II error). Because a Bayesian approach focuses on estimating effects in a single, unified procedure, rather than determining whether each effect is significant via repeated separate hypothesis tests, it avoids the problem of multiple comparisons that arises from repeated testing. Instead, using a hierarchical prior structure controls the risk of spurious overestimation within the model itself. The partial pooling induced by a hierarchical set of priors has the effect of drawing effect estimates closer to one another and toward zero (when the highest-level priors in the model are centered at zero, as is typically the case). The result is that, instead of expanding confidence intervals and leaving effect estimates unchanged, the Bayesian approach produces (appropriately) more conservative effect estimates that do not require subsequent correction to represent statistical precision accurately (Gelman 2012).

IV. MODEL

The full Bayesian model for a factorial experiment consists of a likelihood and a set of hierarchical priors (also known as *random effects* or *shrinkage* priors) for the model's parameters. The likelihood resembles the classical regression model: each level of each factor has a main effect, and there is an interaction effect for each combination of the levels of each combination of factors. In our case, we include up to pairwise interactions between factor levels; in advance of the study, we determined on a theoretical basis that three-way or higher dimension interactions were likely to be very small (Li et al. 2006). The main effect and pairwise interaction effect terms can be equivalently written as either the product of parameters and indicator variables or as sets of main effects and of interaction effects that are indexed by factor and level. (Given the large number of treatment arms, we chose the latter representation for the sake of concision.) Additional covariates are included as additional linear terms as per classical linear regression; in our case we use these terms to control for respondents' demographic characteristics.

[3] These within-factor variance components can be interpreted as a gestalt measure of the importance of each factor on the outcome. These variance parameters are themselves modeled as coming from a common prior that reflects expectations about the overall distribution of effect sizes in the experiment. In the parlance of Bayesian statistics, the parameters of the prior distribution are known as *hyperparameters*, and the priors on the hyperparameters as *hyperpriors*. An astute statistician will note that one could model the parameters of the hyperpriors with priors of their own, and so on. This is unnecessary in practice, and the aspiring Bayesian statistician need not worry about continuing to define even higher-level parameters to govern these hyperpriors. Although partial pooling is not a uniquely Bayesian approach—non-Bayesian mixed models can achieve a similar effect—it would not be possible to estimate the variance components for factors with a very small number of levels in a non-Bayesian setting. Using the hyperprior on the variance components allows us to do this (Gelman and Hill 2016, pp. 498-500).

The experiment defined treatment arms with a set of five factors, described previously. In addition, we made the model scale free; that is, all outcomes were standardized to have a mean of zero and a standard deviation of 1, as were all continuous predictors; binary predictors were left as 0/1. The study analyzed data from respondents in all 72 treatment arms to estimate the following model:

$$y_i = \alpha + \sum_{m \in F} \beta^{(m)}_{j_i^{(m)}} + \sum_{\substack{q,r \in F \\ q \neq r}} \theta^{(q,r)}_{j_i^{(q)}, j_i^{(r)}} + \gamma \cdot X_i + \varepsilon_i.$$

In the equation above, respondents are indexed by i, so that y_i is the outcome of interest for respondent i. The set F is a set of indices representing the five factors in the experiment. For a given factor $m \in F$, the index $j_i^{(m)}$ indicates the level of factor m respondent i receives. The term $\beta_j^{(m)}$ represents the main effect of factor m at level j, and the term $\theta_{k,l}^{(q,r)}$ represents the interaction effect between factor q at level k and factor r at level l. Thus, the term $\beta^{(m)}_{j_i^{(m)}}$ in the likelihood above represents the main effect of factor m on the outcome of respondent i.The vector X_i is a set of additional covariates with effects γ, α is an overall intercept, and ε_i is a respondent-level error term.

The prior distributions for the model's parameters are as follows:

$$\begin{aligned}
&\beta^{(m)} \sim \mathcal{N}\left(0, \tau^{(m)}\right) \\
&\theta^{(q,r)} \sim \mathcal{N}\left(0, \tau^{(q,r)}\right) \\
&\epsilon \sim \mathcal{N}(0, \sigma) \\
&\\
&\tau^{(m)} \sim \mathcal{N}(0, \phi_{\text{main}}) \\
&\tau^{(q,r)} \sim \mathcal{N}(0, \phi_{\text{int}}) \\
&\alpha, \sigma, \gamma, \phi_{\text{int}}, \phi_{\text{main}} \sim \mathcal{N}(0,1)
\end{aligned}$$

Here $\mathcal{N}(0, s)$ indicates either a normal distribution with a mean of zero and standard deviation s, or the corresponding half normal for the standard deviation parameters τ, σ, and ϕ (the term *half normal* refers to a normal distribution truncated below a value of zero, meaning there are no negative values). The first three rows here define priors for the parameters of main interest in the likelihood, while the next two rows define priors for the parameters of these priors. The last row sets the prior for parameters we do not want to model with additional structure or strong prior information, using a distribution that is broad and relatively uninformative on the scale of the model. Some Bayesian statisticians advocate for using Cauchy distributions or even so-called improper infinite uniform priors here, but using a normal distribution provides additional computational stability and does not represent a strong assumption about the parameters. In selecting these priors, we followed previous work (Gelman 2006) and the current recommendations from the Stan Development Team (2017).

Rather than pick a baseline or reference level for each factor in the model, we explicitly include a term $\beta_m^{(m)}$ for every level of each factor in our model. To preserve identifiability of the model, we impose the constraint that the main effects for the levels of each factor must sum to

zero: $\sum_m \beta_m^{(m)} = 0$. The effect of a factor is read off relative to zero (and zero is by definition the mean of the effects for each factor). We also prefer this approach for the sake of interpretation in our results, as no clear baseline category exists for the school information design strategies tested in our experiment. We use analogous contrasts for the interaction terms: we explicitly model an interaction term $\theta_{p,q}^{(p,q)}$ for each combination of levels of each pair of factors and impose the constraint that these effects sum to zero within each pair of factors: $\sum_{p,q} \theta_{p,q}^{(p,q)} = 0$. This choice of contrasts for interaction effects means the expected effect of a given factor level is not, in general, equivalent to the main effect $\beta_m^{(m)}$ of that level read directly from the model. To read off the full effect of a given factor level, we add to the main effect the average of all interaction terms that involve that factor level: the total effect of factor m at level $j^{(m)}$ is given by $\beta_{j^{(m)}}^{(m)} + \sum_{\substack{q \in F \\ q \neq m}} \frac{1}{J^{(q)}} \sum_{j^{(q)}} \theta_{j^{(m)},j^{(q)}}^{(m,q)}$, where $J^{(q)}$ is the number of levels of factor q.

V. POWER ANALYSES FOR BAYESIAN FACTORIAL EXPERIMENTS

Although the Bayesian framework for factorial experiments has several advantages, determining the number of treatment arms that can be tested with a given sample size can be challenging. The statistical precision of our model depends on partial pooling, which in turn depends on how many of the factors (and interactions between factors) prove to be important in affecting the study's outcomes of interest, in the sense that the variance of effect sizes turns out to be larger within some factors or interactions compared with others (Gelman 2005). If only a small number of factors and few interactions are important, there will be more pooling and more precision. On the other hand, if many factors and interactions are important, there will be less pooling and greater uncertainty in the effect estimates. This means there is uncertainty regarding the ultimate precision of a factorial experiment with a given sample size and a given number of treatment arms.

To address this design challenge, we estimate likely values for the minimum detectable effect (MDE) by running simulated repetitions of the experiment before it occurs. We simulate the experiment under each of a range of possible effect sizes and estimate the smallest effect size that allows us to correctly identify favorable treatment arms with at least 80 percent probability. Our primary analysis seeks to identify which school display (treatment arm) is most suitable for a given outcome by examining paired comparisons of arms. For this analysis, we define the effect of an arm to be the average outcome value of that arm (as given by taking the sum of the regression coefficients for the given arm) and consider the minimum detectable difference between any two arms of the study. Researchers familiar with factorial experiments may be more used to considering the main effects of each factor and considering only the minimum detectable difference between the effects of different levels of the same factor. In a typical frequentist setting, a researcher would model only these main effects and compare regression coefficients, but the borrowing of strength induced in our Bayesian setting allows us to model interactions as well. As noted above, the effect of a given factor level in our model is calculated as the main effect regression coefficient plus the average of those interaction effects that include the given factor level. In the power analyses discussed below, we examine MDEs for these main effects in addition to the paired comparison of treatment arms that served as the primary contrast of interest in our study.

To carry out simulated repetitions of the experiment for a given sample size and number of treatment arms, we first randomly draw treatment and interaction effects from centered normal distributions, whose variance parameters are themselves randomly drawn from a half-Cauchy distribution with fixed-scale parameter. For generating simulated data, the Cauchy distribution is a good choice to model situations with mostly small effects, but a few large ones (Gelman 2006). This differs from the final model used for analysis of the simulated data, which used half-normal priors. Changing to half-normal priors for analysis provided our model software with greater computational stability and reflected evolving consensus in the Bayesian community (Stan Development Team 2017); this change does not noticeably impact the model output.

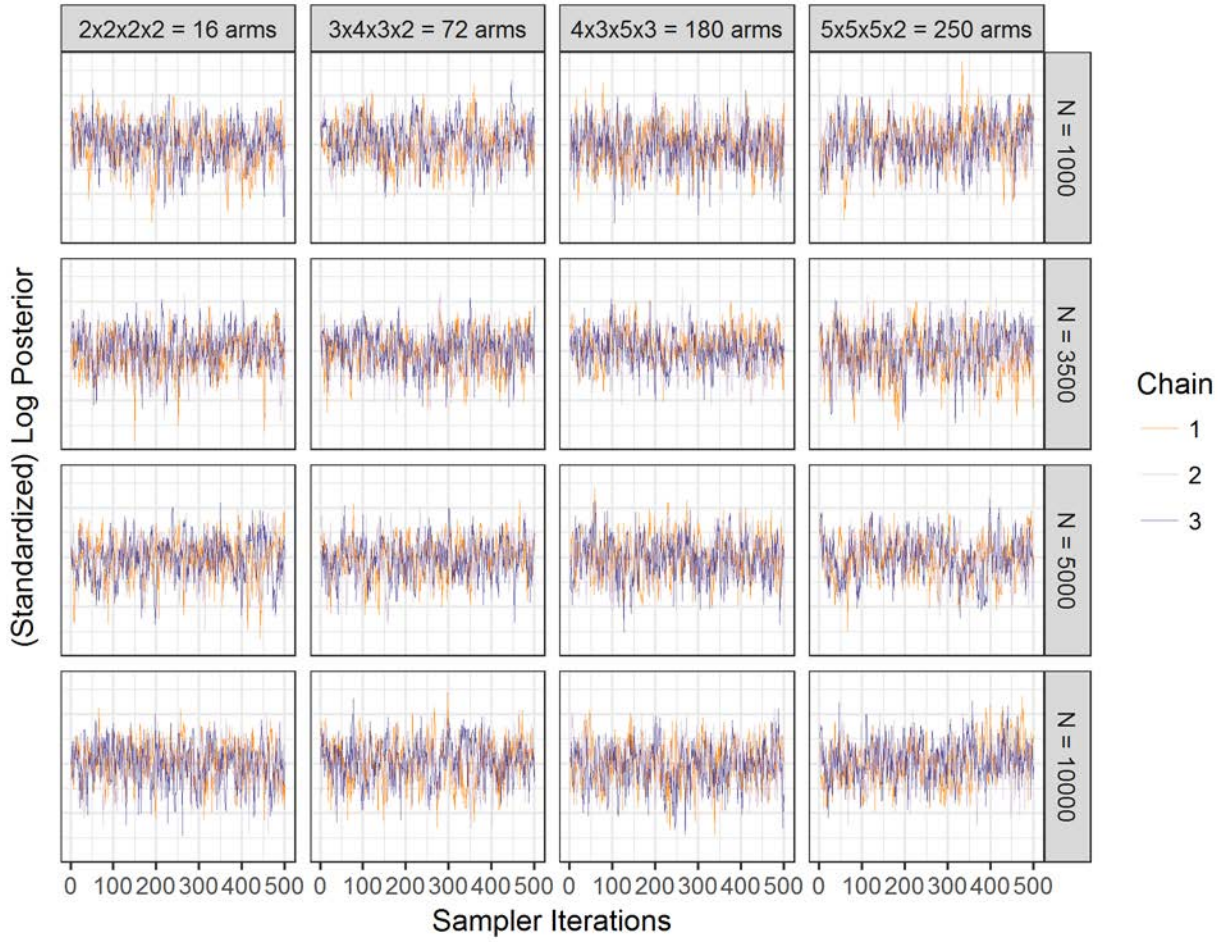
We use the generated treatment and interaction effects in each simulation to calculate the so-called true mean effect of being in each combination of factor levels. We then transform these mean effects to be centered at zero and scaled such that the maximum effect has a fixed and pre-specified size. We simulate individual participants in the experiment by taking a number of normal draws from distributions centered at the effect of each treatment arm, with predetermined noise variance. In our experiment, we set our half-Cauchy scale parameter to 5, the difference between the maximum and mean treatment effect to 0.25 (as per findings in Jacobsen et al. 2014), and the variance of individual noise to be 0.88, with this last value chosen such that 12 percent of the variance in outcomes is explained by observed demographic characteristics (as per findings in Tuttle et al. 2013).

We fit the Bayesian model to the simulated data using Stan (Carpenter et al. 2017). We use this model to find the posterior probability of correctly identifying which of the two treatment arms in a given pair has the greater effect. Those pairs of treatment arms for which this probability is over a certain threshold are considered significant findings. We then fit a logistic regression to predict this binary significance from the true effect difference in each pair of treatment arms. Based on this logistic regression, we consider the MDE of this experiment to be the smallest difference in effect size with at least an 80 percent chance of finding it significant in the correct direction by the Bayesian model. For highly underpowered simulations, such as those when no significant differences between arms are found, we treat the MDE as effectively infinite in our results. The choice to use a binary significance outcome in this regression was motivated by a desire to help explain our calculations in familiar terms to stakeholders who were unaccustomed to the language of Bayesian statistics. However, there is no theoretical reason why the posterior probability cannot be used directly as the outcome in the regression. In our experiment, we set the threshold for significance at 0.975 to correspond to a two-sided *p*-value at the 95-percent confidence level, but experimenters may wish to explore other threshold values or even consider dispensing with the intermediate significance calculation altogether.

Although Stan's implementation of the Hamiltonian Monte-Carlo algorithm is generally stable and effective, when using this type of Markov Chain Monte Carlo (MCMC) method for Bayesian inference it is important to verify that the sampler has converged to the posterior distribution with sufficient "mixing" of the chains to yield an adequate number of effectively independent posterior draws. For each model, we ran three chains with 500 iterations of burn-in and estimated posterior distributions from a subsequent 500 iterations of the chain. None of the parameters of the model had a Gelman-Rubin potential scale reduction factor ($\hat{R}$) above 1.1, a key statistic consistent with the sampler having converged to the true posterior. Furthermore, none of our effects for factors, interactions, or arms was estimated from less than 100 effectively

independent posterior draws, and visual inspection of trace plots of the log-posterior and key model parameters were also consistent with the chains of the sampler mixing well. Figure 2 shows examples of these trace plots. Finally, no iterations of the sampler encountered divergent transitions or exceeded the maximum tree depth.

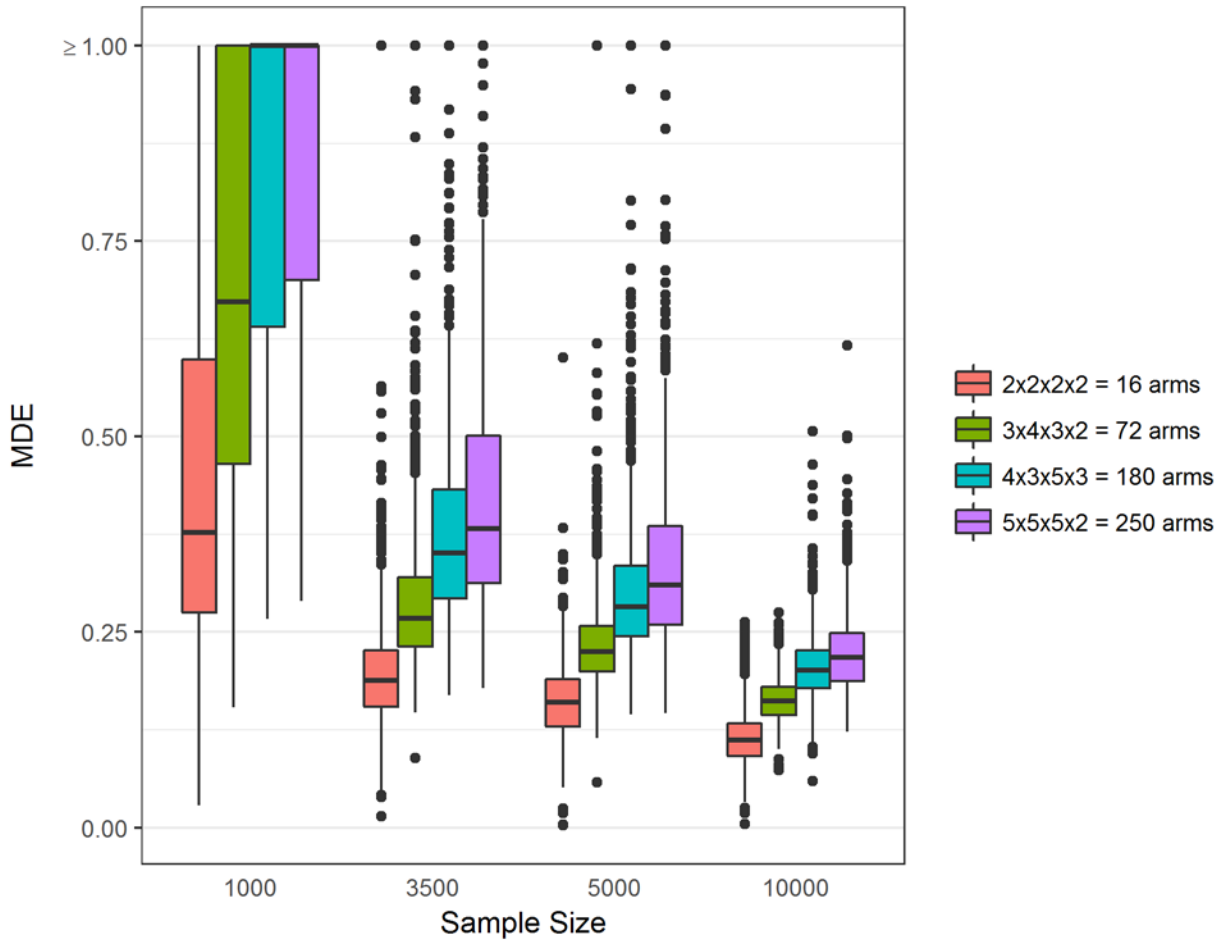
Figure 2. Traceplots for representative Stan output



Note: This figure shows the trace plots for the log posterior (on a standardized scale) for each sample size and experimental configuration of a single simulation randomly chosen as a representative. The chains do not remain stationary or monotonic for any long period, a sign that they are well mixed.

Repeating this simulation process a number of times gives a distribution of MDEs for each candidate sample size and number of treatment arms. These distributions are depicted in the box-and-whiskers plot in Figure 3, which shows the quartiles and outliers for each set of study sizes (four different sample sizes, four different numbers of treatment arms). These distributions can be used to estimate our uncertainty about the MDE. Although it is not common practice to estimate this uncertainty in non-Bayesian power calculations, such uncertainty is always present; the precision of an experiment is never completely certain a priori. These simulations merely make this fact explicit and allow experimenters to account for the uncertainty about the MDE when finalizing their choice of sample size and the number of treatment arms.

Figure 3. Distribution of MDE estimates, by sample size and number of treatment arms



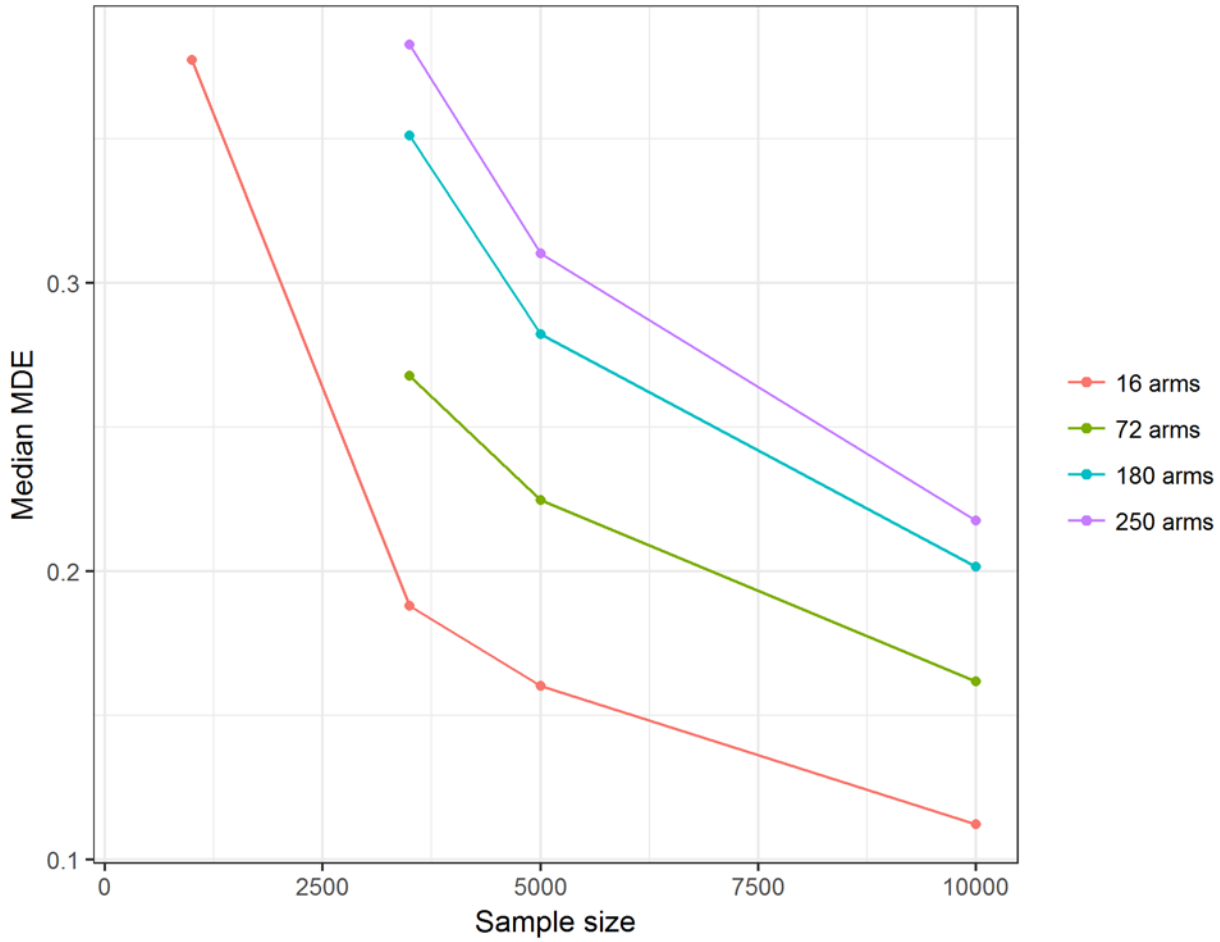
Note: This figure illustrates the quartiles for the distribution of possible MDEs for each study design as a box-and-whiskers plot. The figure shows MDE results for four different factorial designs (16 arms, 72, arms, 180 arms, and 250 arms), across four different sample sizes. Each of the 16 box plots in the figure summarizes the results of 1,000 simulations, and outlier MDE values are shown as individual data points. The figure shows that the median MDE rises with the number of arms in the experiment but declines with larger sample sizes. In addition, there is less around the median MDE estimate as sample sizes increase (that pattern is evident for all study sizes shown, but is less noticeable for studies with fewer treatment arms).

We use the median of these distributions as a point estimate for the MDE; doing so allows us to limit the influence of the effectively infinite MDEs that arise from the occasional underpowered simulation. However, if more than one-third of the simulations for a given sample size and number of treatment arms have power issues that prevent calculating a finite MDE, we treat our results as though the experiment cannot be run at this size and omit the results. Consequently, several points are missing from subsequent charts of our power simulation results.

Figure 4 illustrates the relationship between sample size and MDE as a line plot, which shows the unsurprising result that increased sample size corresponds to lower MDE, albeit with diminishing returns for larger sample sizes. Charts like this allow researchers to carefully select a

number of treatment arms and sample size that are compatible with their desired MDE. The results also show that, for a fixed sample size, there are diminishing marginal costs (in terms of the MDE) of increasing the number of study arms. This is more pronounced for studies with larger sample sizes. That is, large sample sizes are doubly valuable for a researcher wishing to perform a many-armed experiment, as they lead to lower overall MDEs and reduce the marginal loss of precision from adding more treatment arms.

Figure 4. Median MDE estimate versus sample size, by number of treatment arms



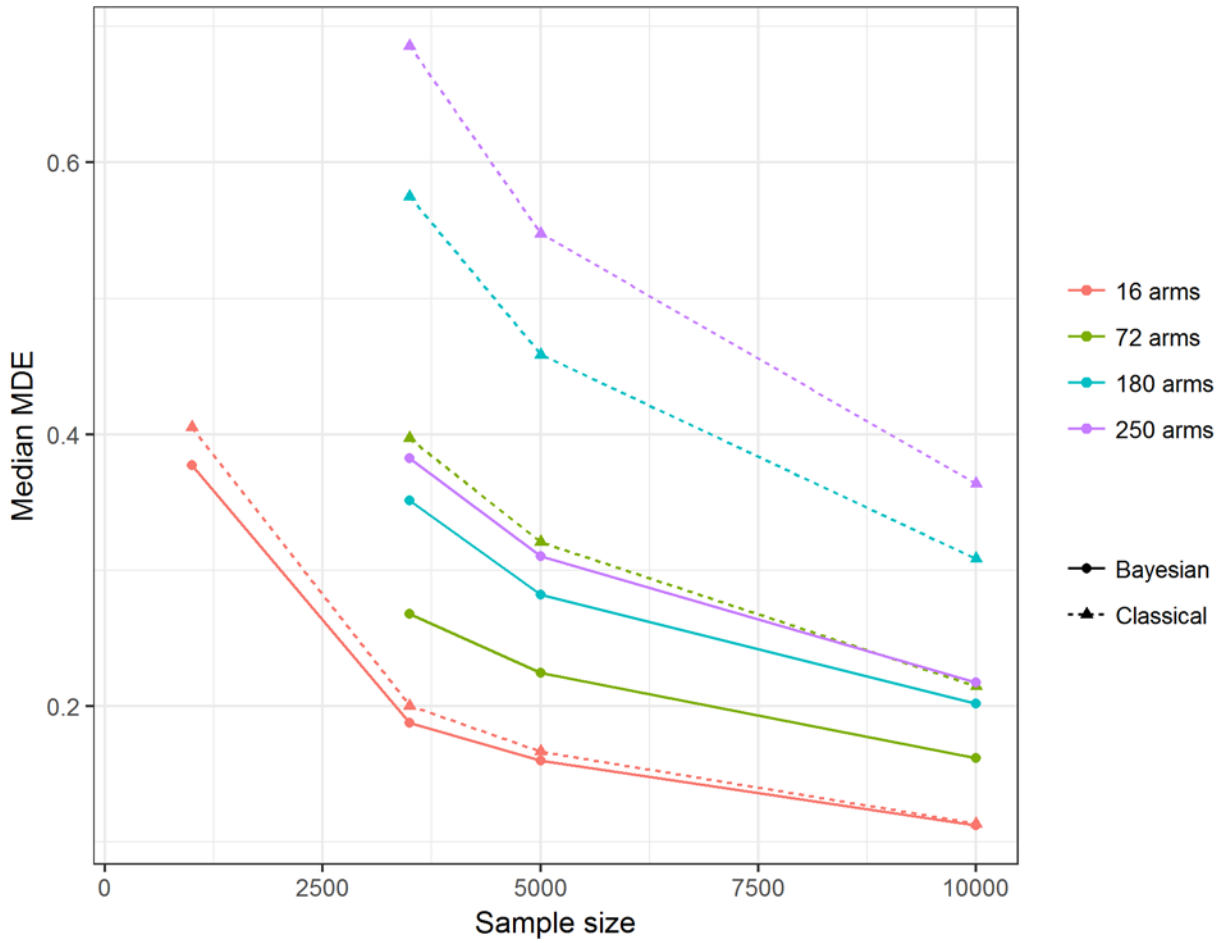
Note: The figure summarizes the expected MDE for four different factorial designs (16 arms, 72, arms, 180 arms, and 250 arms), across four different sample sizes. For each design, the data points plotted in the figure represent the median MDE across 1,000 simulations. If more than one-third of the simulations for a given sample size and number of treatment arms have power issues that prevent the calculation of a finite MDE, we omitted the results from the figure. This figure shows that as sample sizes increase (horizontal axis) there is a decline in the median of the distribution of estimated MDEs (vertical axis) for each study design.

To compare these results with the precision of a non-Bayesian approach, we performed simulations to estimate the MDE of the experiment in a classical setting. We performed these simulations using the same process described previously, with only two differences. First and most importantly, instead fitting a Bayesian model using hierarchical priors to induce partial

pooling, we fit a classical, frequentist regression model with the same covariates. Second, instead of determining significance using posterior probabilities (which are a feature of a Bayesian model), we used *p*-values. We adjusted these *p*-values with the Benjamini-Hochberg correction for false discoveries in the presence of multiple comparisons before checking for significance at the 0.05 level.

These simulations show that the Bayesian factorial design provides substantial gains in precision over traditional methods. Figure 5 shows the median MDEs for both sets of simulations (Bayesian and classical) under a range of study sizes. The Bayesian MDEs are similar to

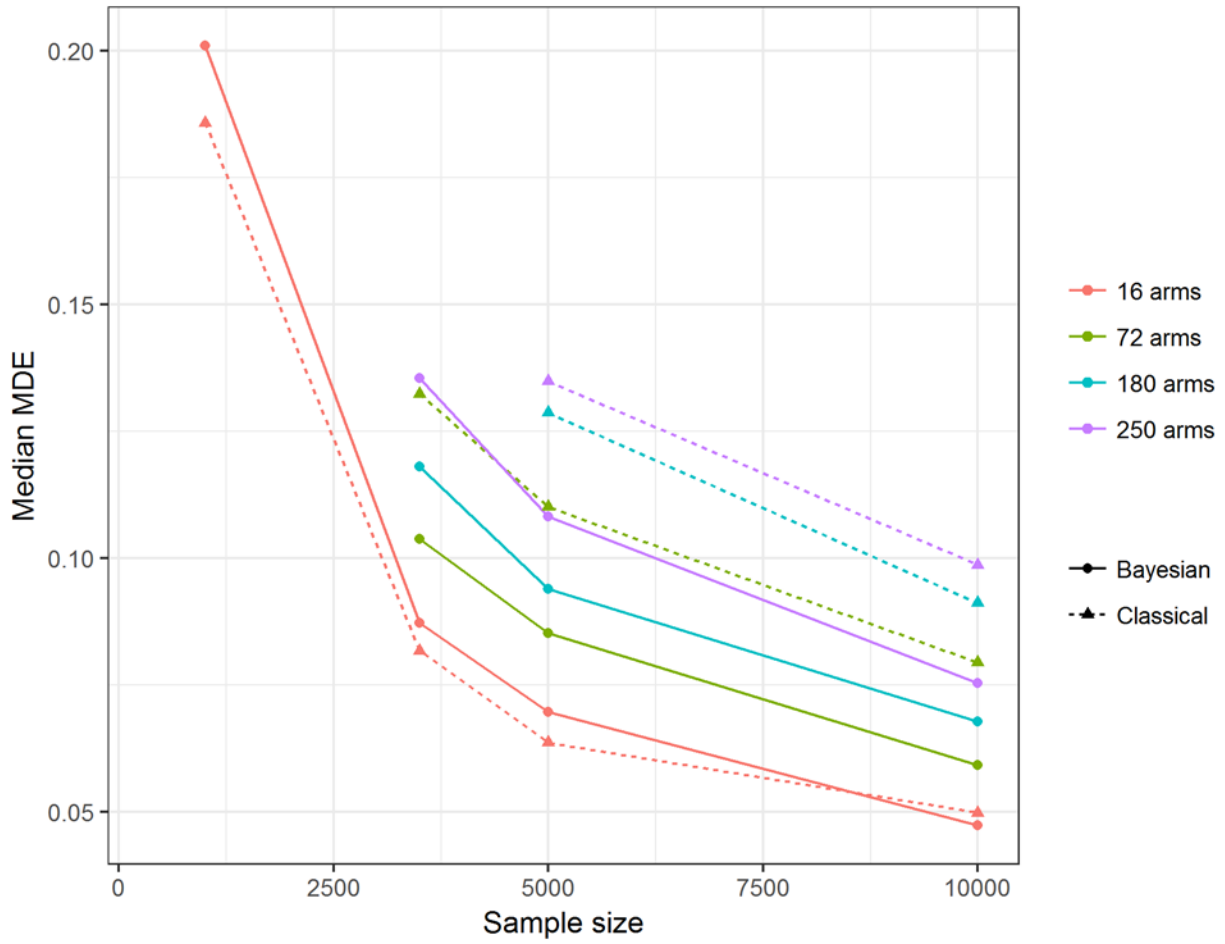
Figure 5. Comparison of Bayesian and classical median MDE estimates, paired comparison of arms



Note: Comparing Bayesian and classical designs, this figure shows the relationship between sample size (horizontal axis) and the median of the distribution of estimated MDEs (vertical axis) for each study design, grouped into lines by the number of treatment arms. In all cases, the MDE for the Bayesian design is lower than that for the corresponding classical design, but the differences are larger for factorial designs with larger numbers of treatment arms. Certain combinations of sample size and number of treatment arms (such as the 16-arm experiment with 1,000 respondents) are missing from this plot because more than one-third of simulations did not produce a finite MDE estimate, a sign of extremely low power.

classical estimates for studies with few treatment arms, but considerably better for complex studies. For example, to test 72 treatment arms for a given MDE, a classical experiment requires roughly twice the sample size as a Bayesian experiment. Figure 6 shows the same summary information for the so-called main effects MDE calculations: the relative gains in precision from the Bayesian approach remain similar when considering main effects.

Figure 6. Comparison of Bayesian and classical median MDE estimates, main effects



Note: Comparing Bayesian and classical designs, this figure shows the relationship between sample size (horizontal axis) and the median of the distribution of estimated MDEs (vertical axis) for each study design, grouped into lines by the number of treatment arms. The relative performance between Bayesian and classical designs follows a similar pattern to the pairwise comparison results in Figure 5. As with Figure 5, certain combinations of sample size and number of treatment arms (such as the 16-arm experiment with 1,000 respondents) are missing from this plot because more than one-third of simulations did not produce a finite MDE estimate, a sign of extremely low power.

Based on the results of these power calculations and discussions with the broader research team involved with the study, we elected to proceed with a design using 72 treatment arms and 3,500 study participants. This was sufficient to provide us with a median MDE of approximately 0.25 standard deviations in a comparison of any two treatment arms, which was consistent with

the magnitude of effects that had been observed in other studies of school information displays and how they affect parents (for example, Jacobsen et al. 2014).

VI. DISCUSSION: APPLYING THE DESIGN

Although a factorial design offers clear and substantial benefits, a researcher intending to embark on such an experiment must plan for the greater logistical complexity that comes with running a large study with many treatment arms. This challenge does not inherently arise from Bayesian methods—any large factorial experiment would need to exercise such caution—but the ability to test many more treatment arms using a given sample size makes such complex experiments more feasible. In the case of the education study examined here, we managed the entire experiment and all of its interventions in the context of a single web-based survey. Managing the study in this way allowed the research team to use online tools to randomly assign participants to treatment arms and track their progress. This made administering the experiment substantially easier—the experiment's logistical complexity was mostly driven by carefully designing the 72 information displays and ensuring that the study's random assignment and survey procedures operated smoothly. In the context of a field experiment testing variations in a policy or program, issues of intervention design, intervention implementation, contamination across treatment arms, and differential attrition would make managing such a large number of interventions and treatment arms considerably harder.

There are several other challenges that come with the use of a Bayesian approach. First, a Bayesian experiment requires more upfront work to design: power calculations must consider a wider range of study designs, and selecting an appropriate model requires careful review of existing literature and precise a priori reasoning about the experiment. Bayesian models force researchers to make explicit assumptions about their experiment in the form of priors, and although non-Bayesian models implicitly make their own assumptions (often implausible ones), many researchers are discouraged by the task of setting a good prior. More specifically, to perform sample size computations for their own Bayesian factorial experiment, researchers would need to follow these steps:

1. **Hypothesize treatment effects.** To set up this power calculation, researchers need to make assumptions about the potential size and distribution of true treatment effects for the selected set of intervention factors (and specify hypothetical distributions for the other parameters in the study's analytical model).
2. **Design the factorial study.** The next step in the power analysis is to specify the design of the study and the candidate configuration (or configurations, if the power analysis seeks to compare multiple design options) of factors and factor levels.
3. **Calculate power using many simulated runs of the experiment.** This involves repeatedly simulating data from the hypothetical distribution specified in step 1 (our power analyses used 1,000 simulations) and then fitting a multilevel model to each simulated data set. The power analysis returns the proportion of the simulations where a given effect size for one of the contrasts of interest passes a selected posterior probability threshold for detection. By examining the distribution of results from these simulations, researchers can select the necessary sample size to achieve the desired MDE for their experiment.

In addition to the added steps involved with power calculations, Bayesian methods do entail other potential challenges. Bayesian inference is somewhat more computationally intensive than classical methods. Until recently, the time and resources required for Bayesian modeling made it impractical for large and complex experiments. However, in recent years new software has made this approach much more feasible (Carpenter et al. 2017). Finally, the relative newness of Bayesian methods in the policy sphere may make explaining the study's methods and results more difficult to policymakers who are accustomed to conventional approaches that use statistical significance. For example, the U.S. Department of Education's What Works Clearinghouse has yet to develop explicit standards for assessing the modeling decisions in Bayesian studies or guidance for how impact findings generated using Bayesian models should be compared with results estimated using classical models. In a high-stakes evaluation, researchers should carefully consider how the decision to apply Bayesian methods will impact their experimental process and be viewed among decision makers.

That said, we believe that the type of large factorial experiment enabled by Bayesian modeling has the potential to reveal readily applicable insights and help experimenters uncover optimal combinations of tested practices for a wide range of practitioners and policymakers. With the benefits of this analytical approach, the primary barrier to conducting complex experiments should be the logistics of managing many treatment arms, rather than concerns over sample size or the precision of effect estimates. We look forward to seeing factorial experiments applied in a broader range of policy areas and contexts.

REFERENCES

Benjamini, Y., and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B (Statistical Methodology),* vol. 57, no. 1, 1995, pp. 289–300.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software,* vol. 76, no. 1, 2017. doi:10.18637/jss.v076.i01.

Chakraborty, B., L. M. Collins, V. Strecher, and S. A. Murphy. "Developing Multicomponent Interventions Using Fractional Factorial Designs." *Statistics in Medicine,* vol. 28, 2009, pp. 2687–2708.

Collins, L. M., J. J. Dziak, and R. Li. "Design of Experiments with Multiple Independent Variables: A Resource Management Perspective on Complete and Reduced Factorial Designs." *Psychological Methods,* vol. 14, no. 3, 2009, pp. 202–224.

Cox, D. R. *Planning of Experiments.* New York: John Wiley and Sons, 1958.

Dziak, J. J., I. Nahum-Shani, and L. M. Collins. "Multilevel Factorial Experiments for Developing Behavioral Interventions: Power, Sample Size, and Resource Considerations." *Psychological Methods,* vol. 17, no. 2, 2012, pp. 153–175. doi:10.1037/a0026972.

Gelman, A. "Analysis of Variance: Why It Is More Important Than Ever." *The Annals of Statistics,* vol. 33, no. 1, 2005, pp. 1–53. doi:10.1214/009053604000001048.

Gelman, A. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis,* vol. 1, no. 3, 2006, pp. 515–533.

Gelman, A. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness,* vol. 5, no. 2, 2012, pp. 189–211. doi:10.1080/19345747.2011.618213.

Gelman, A., and J. Carlin. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science,* vol. 9, no. 6, 2014, pp. 641–651. doi:10.1177/1745691614551642.

Gelman, A., and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press, 2007.

Gelman, A., J. B. Carlin, H. S. Sterm, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis* (3rd ed.). Boca Raton: CRC Press, 2015.

Jacobsen, R., J. Snyder, and A. Saultz. "Information or Shaping Public Opinion? The Influence of School Accountability Data Format on Public Perceptions of School Quality." *American Journal of Education,* vol. 121, no. 1, 2014, pp. 1–27.

Li, Xiang, Nandan Sudarsanam, and Daniel D. Frey. "Regularities in Data from Factorial Experiments." *Complexity,* vol. 11.5, 2006, pp. 32–45.

Nair, V., V. Strecher, A. Fagerlin, P. Ubel, K. Resnicow, S. A. Murphy, R. Little, B. Chakraborty, and A. Zhang. "Screening Experiments and the Use of Fractional Factorial Designs in Behavioral Intervention Research." *American Journal of Public Health,* vol. 98, 2008, pp. 1354–1359.

Nichols-Barrer, Ira, Alyson Burnett, Steven Glazerman, Jon Valant, and Jesse Chandler, "Parent Information and School Choice Evaluation: Design Report." Final report submitted to the U.S. Department of Education, Institute of Education Sciences. Washington, DC: Mathematica Policy Research, May 2016.

Schochet, P. Z. *Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions.* Washington, DC: National Center for Education Evaluation and Regional Assistance, 2008.

Stan Development Team. "Prior Choice Recommendations."Available at https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. Accessed June 4, 2018.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP Middle Schools: Impacts on achievent and other outcomes.* Washington, DC: Mathematica Policy Research.

Waller, R., and D. Duncan. "A Bayes Rule for the Symmetric Multiple Comparisons Problem." *Journal of the American Statistical Association,* vol. 64, no. 328, 1969, pp. 1484–1503.

ABOUT THE SERIES

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact Ira Nichols-Barrer, Senior Researcher, at inichols-barrer@mathematica-mpr.com.

Suggested citation: Kassler, Daniel, Ira Nichols-Barrer, and Mariel Finucane, "Beyond 'Treatment versus Control': How Bayesian Analysis Makes Factorial Experiments Feasible in Education Research." Working Paper 61. Cambridge, MA: Mathematica Policy Research, June 2018.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ **ANN ARBOR, MI** ■ **CAMBRIDGE, MA** ■ **CHICAGO, IL** ■ **OAKLAND, CA** ■ **SEATTLE, WA** ■ **TUCSON, AZ** ■ **WASHINGTON, DC** ■ **WOODLAWN,** MD